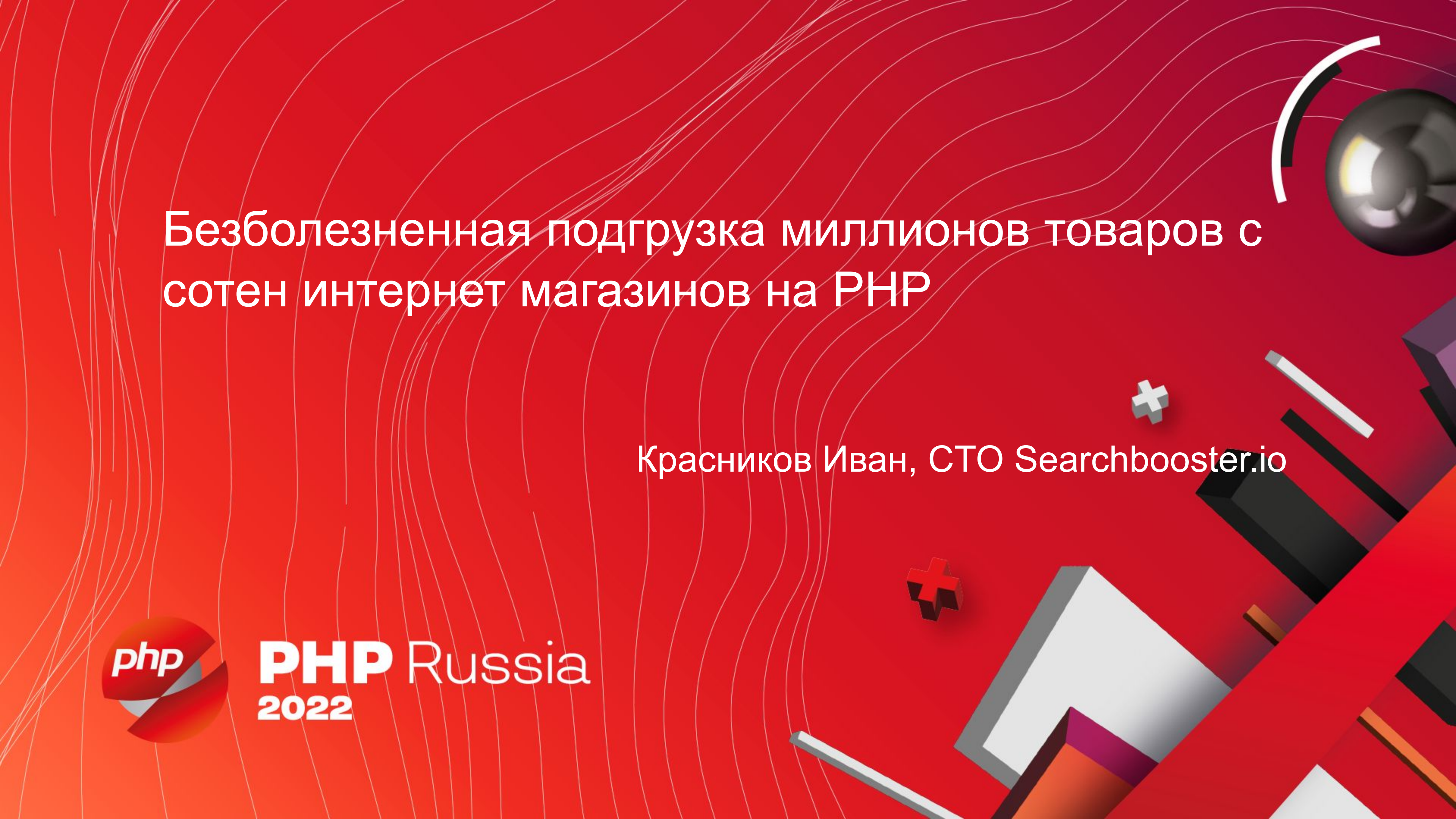
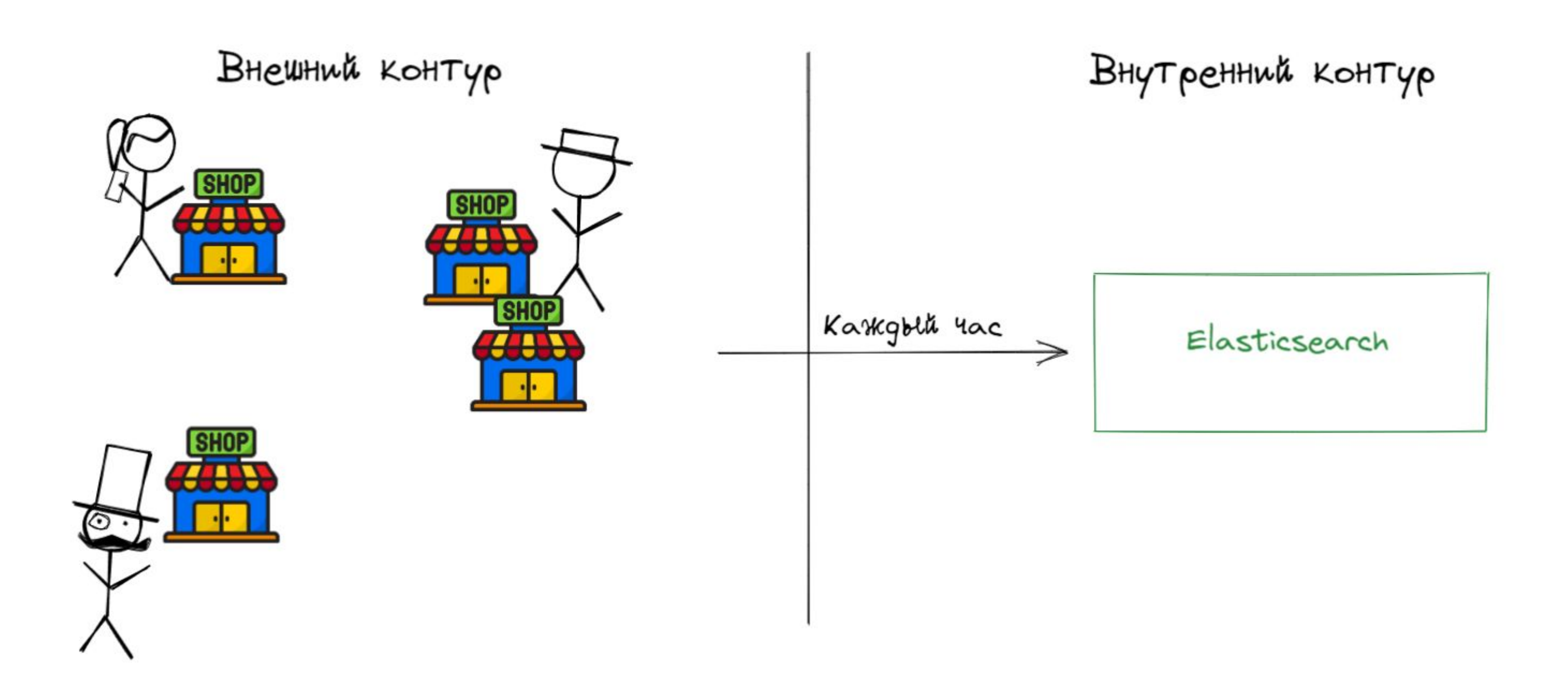
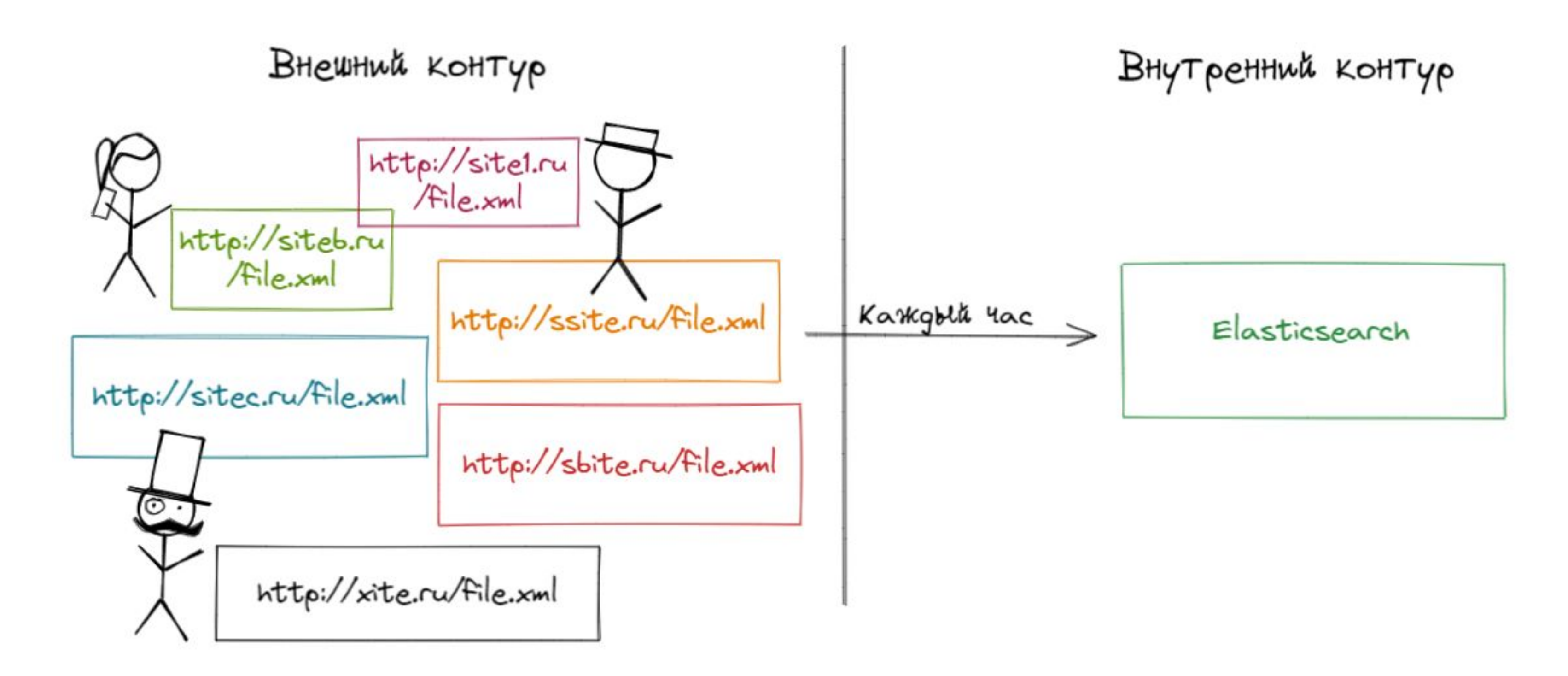

Безболезненная подгрузка миллионов товаров с сотен интернет магазинов на PHP
Красников Иван, CTO Searchbooster.io
php
PHP Russia
2022

Задача



Внешний контур
Внутренний контур
SHOP
SHOP
SHOP
SHOP
Каждый час
Elasticsearch

Задача


Внешний контур
Внутренний контур
http://site1.ru
/file.xml
http://siteb.ru
/file.xml
http://ssite.ru/file.xml
http://sitec.ru/file.xml
http://sbite.ru/file.xml
http://xite.ru/file.xml
Каждый час
Elasticsearch

Общий вид

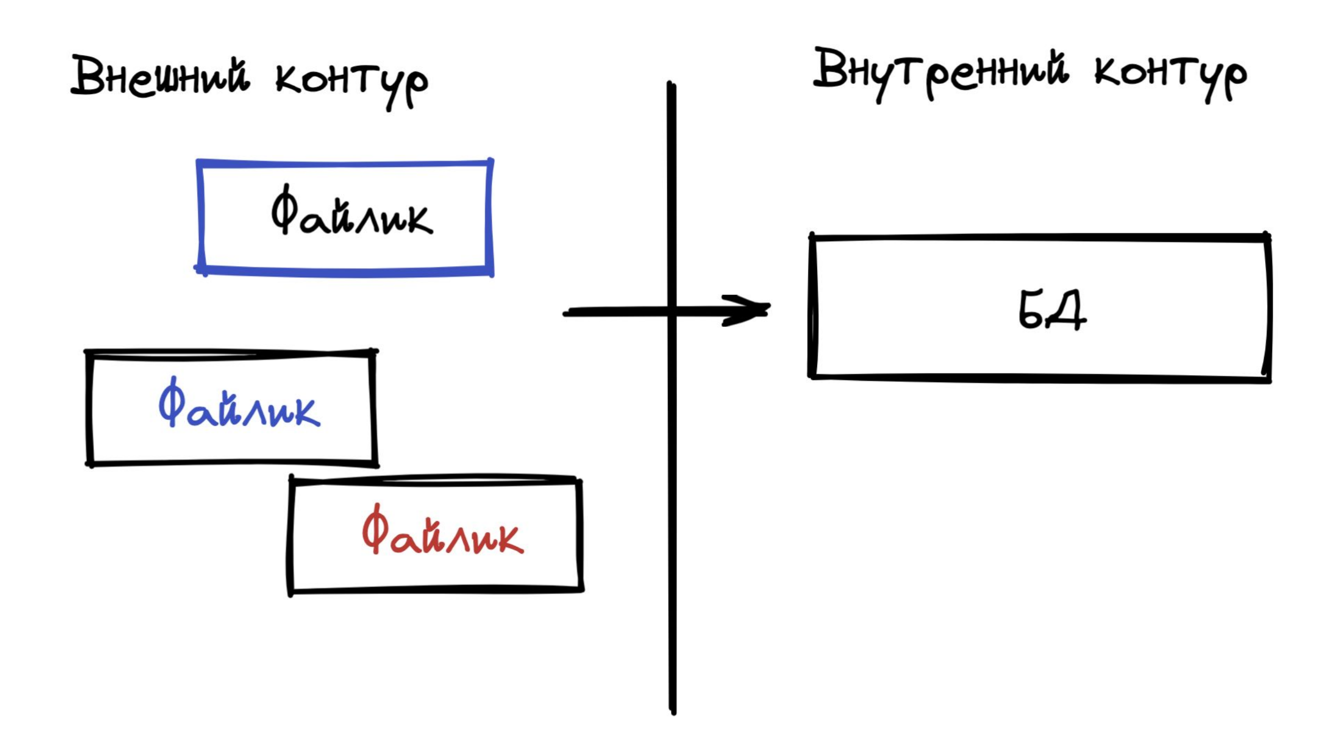
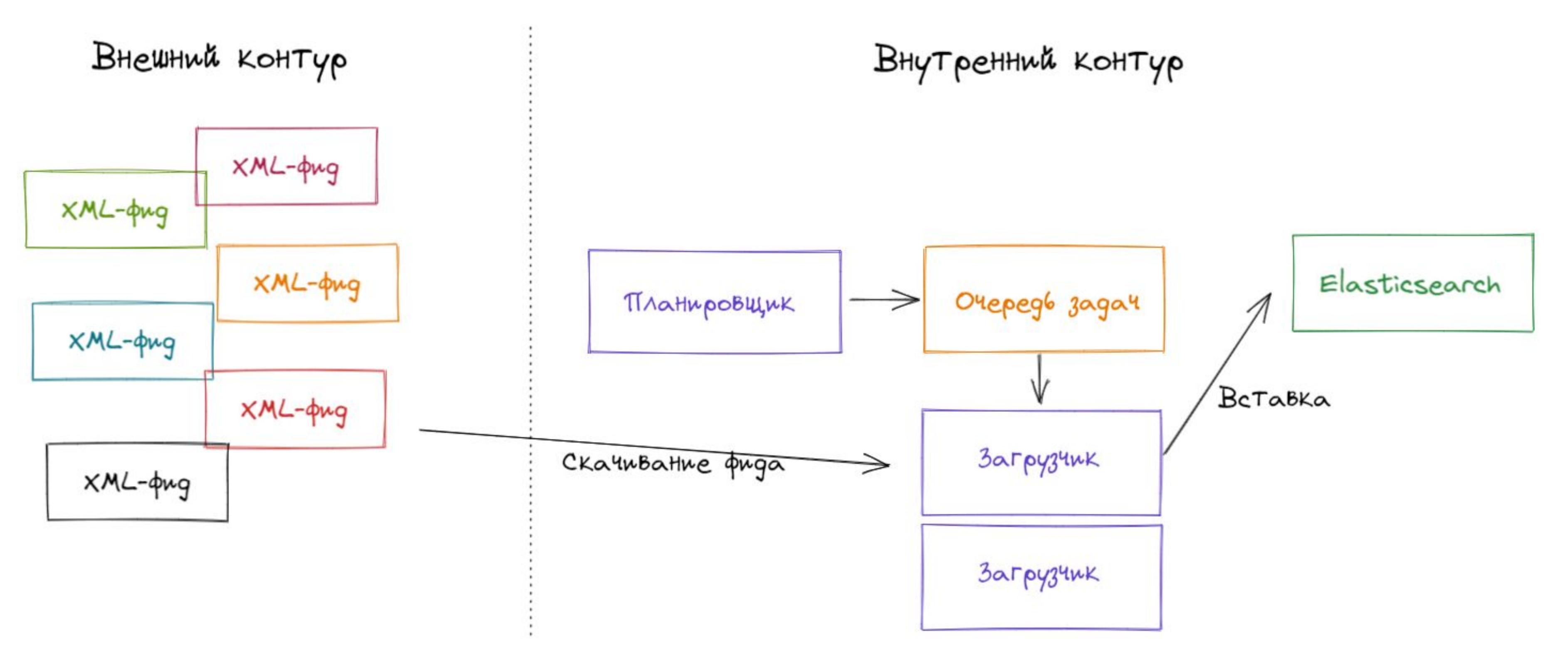


Схема работы



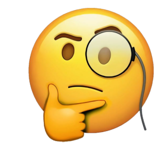


Хотелка

SUPPORT:
Клиенты часто спрашивают, почему их фид не индексировался, а там то фид невалидный, то их сервер лежал. Время тратим на выяснение

А может быть, будем показывать клиентам куски логов с причинами неудачных загрузок?

DEV:



Логи и алерты: публичные

Показываем пользователю лог неудачных загрузок

История загрузок фида

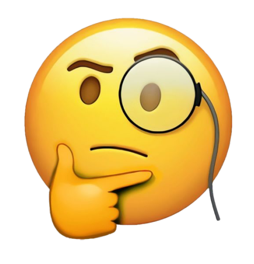
Всего: 15

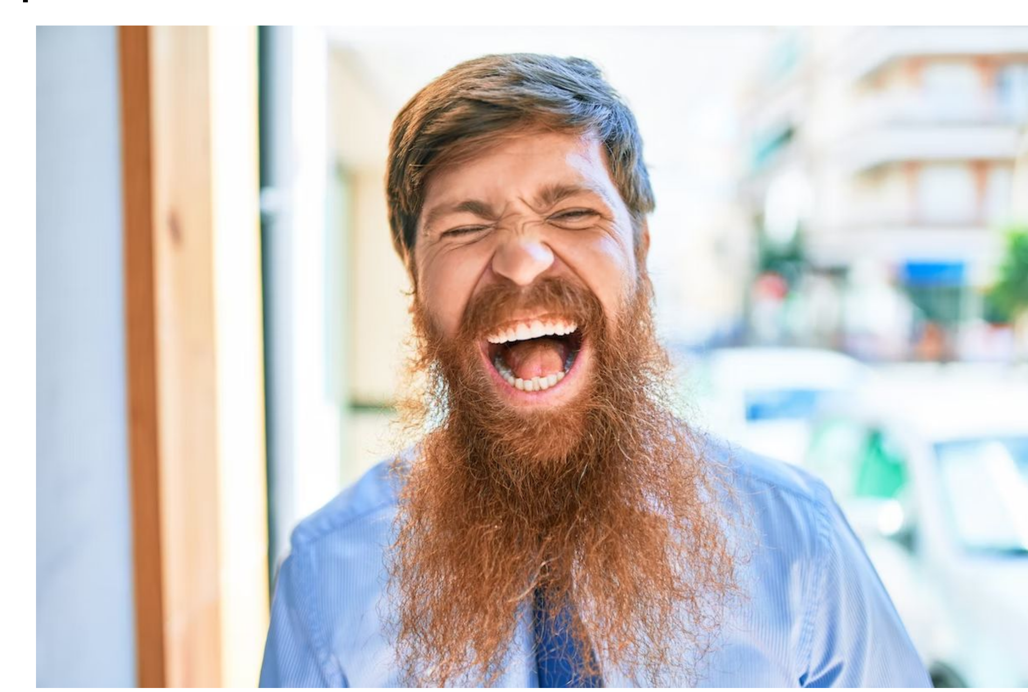
Название	Параметры	Дата
source_not_resolved ()	{ "url": "http://blog.nenado.info", "error": "Could not resolve host: blog.nenado.info", "http_code": 0 }	15.10.2022 11:17:37
source_not_resolved ()	{ "url": "http://blog.nenado.info", "error": "Could not resolve host: blog.nenado.info", "http_code": 0 }	15.10.2022 11:06:36
source_not_resolved ()	{ "url": "http://blog.nenado.info", "error": "Could not resolve host: blog.nenado.info", "http_code": 0 }	15.10.2022 10:55:36

Проблема с загрузкой фида

SUPPORT:
Спасибо за логи!
Тут Клиент спрашивает: “А что за странная ошибка в логе?”

DEV:





Логи и алерты: публичные

История загрузок фида

Всего: 15

Название	Параметры	Дата
source_load_error	cURL error 7: Failed to connect to rc1a-***.mdb.yandexcloud.net port 8123: Connection refused (see https://curl.haxx.se/libcurl/c/libcurl-errors.html) for http://rc1a-***.mdb.yandexcloud.net:8123?wait_end_of_query=1&database=searchbooster&user=user&password=password	15.10.2022 10:26:36

Вывод 1

Показывать пользователю ошибки – хорошая фича, но только заранее предусмотренные ошибки.

Проблема с загрузкой фида

PM:

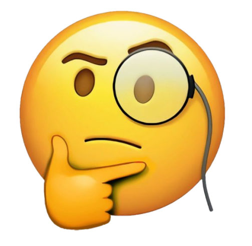
Не грузится фид! Ошибка:
не удается скачать файл.

DEV:

А этот фид открывается
 в браузере?

PM: Да

DEV:



Дебажим!

```
<?php
$url = "https://....xml";

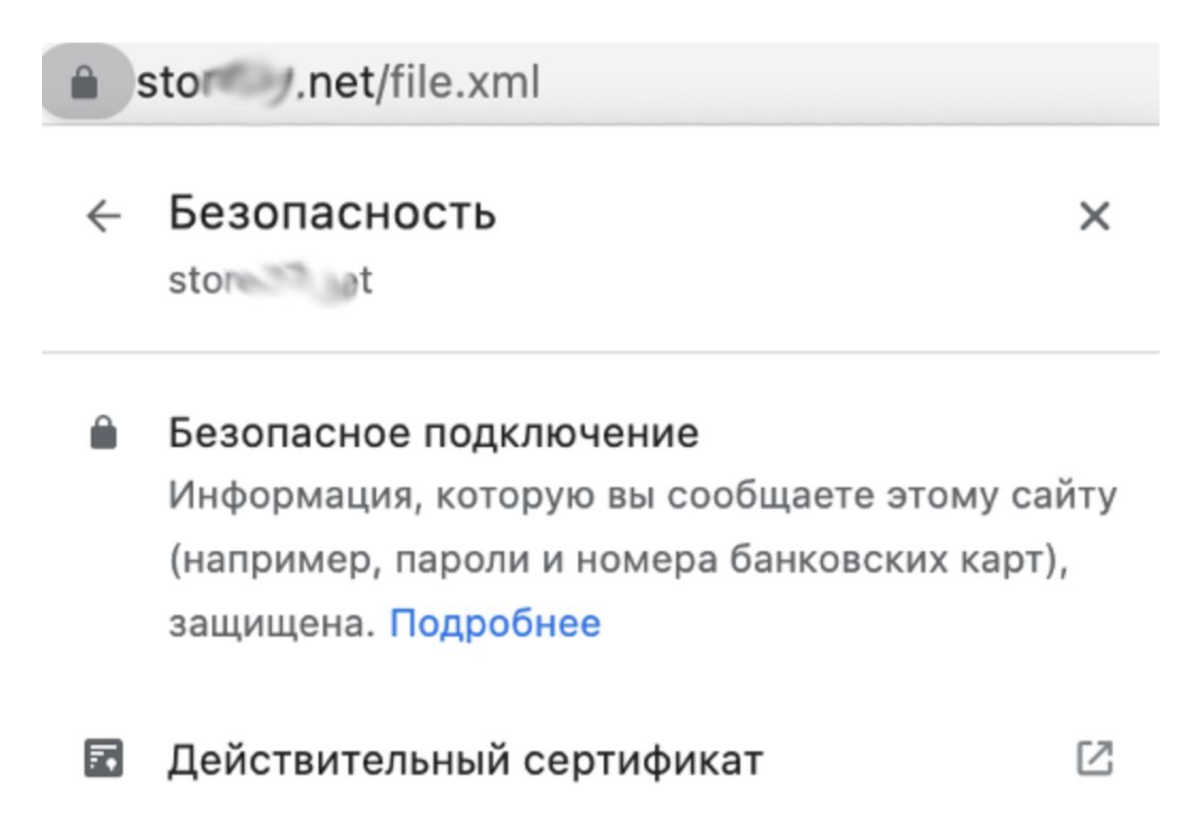
$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
$fp = fopen("file.xml", "w");
curl_setopt($ch, CURLOPT_FILE, $fp);


if (curl_exec($ch) === false) {
    echo "Ошибка curl: " . curl_error($ch);
} else {
    echo "Операция завершена без каких-либо ошибок";
}
curl_close($ch);
fclose($fp);
```

Проблема с загрузкой https-фида

```
php % php test.php
Ошибка curl: SSL certificate problem: unable to get local
issuer certificate
```

Загрузка: проблема SSL



```php
<?php

$url = "https://....xml";

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
$fp = fopen("file.xml", "w");
curl_setopt($ch, CURLOPT_FILE, $fp);
curl_setopt($ch, CURLOPT_VERBOSE, true);

if (curl_exec($ch) === false) {
    echo 'Ошибка curl: ' . curl_error($ch);
} else {
    echo 'Операция завершена без каких-либо ошибок';
}
curl_close($ch);
fclose($fp);
```

Загрузка: проблема SSL

```
* Connected to another.shop (1.1.1.1) port 443 (#0)
* ALPN: offers h2
* ALPN: offers http/1.1
* CAfile: /usr/local/etc/openssl@1.1/cert.pem
* CApath: none
* SSL certificate problem: unable to get local Issuer
certificate
* Closing connection
```

Загрузка: проблема SSL: файл cert.pem

```
GlobalSign Root CA
==================
-----BEGIN CERTIFICATE-----
MIIDdTCCAl2gAwIBAgILBAAAAAABFUtaw5QwDQYJKoZIhvcNAQEFBQAwVzELMAkGA1UEBhMCQkUx
GTAXBgNVBAoTEEdsb2JhbFNpZ24gbnYtc2ExEDAOBgNVBAsTB1Jvb3QgQ0ExGzAZBgNVBAMTEkds
b2JhbFNpZ24gUm9vdCBDQTAeFw05ODA5MDExMjAwMDBaFw0yODAxMjgxMjAwMDBaMFcxCzAJBgNV
BAYTAkJFMRkwFwYDVQQKExBHbG9iYWxTaWduIG52LXNhMRAwDgYDVQQLEwdSb290IENBMRswGQYD
VQQDExJHbG9iYWxTaWduIFJvb3QgQ0EwggEiMA0GCSqGSIb3DQEBAQUAA4IBDwAwggEKAoIBAQDa
DuaZjc6j40+Kfvvxi4Mla+pIH/EqsLmVEQS98GPR4mdmzxzdzxtIK+6NiY6arymAZavpxy0Sy6sc
THAHoT0KMM0VjU/43dSMUBUc71DuxC73/OlS8pF94G3VNTCOXkNz8kHp1Wrjsok6Vjk4bwY8iGlb
Kk3Fp1S4bInMm/k8yuX9ifUSPJJ4ltbcdG6TRGHRjcdGsnUOhugZitVtbNV4FpWi6cgKOOvyJBNP
c1STE4U6G7weNLWLBYy5d4ux2x8gkasJU26Qzns3dLlwR5EiUWMWea6xrkEmCMgZK9FGqkjWZCrX
gzT/LCrBbBlDSgeF59N89iFo7+ryUp9/k5DPAgMBAAGjQjBAMA4GA1UdDwEB/wQEAwIBBjAPBgNV
HRMBAf8EBTADAQH/MB0GA1UdDgQWBBRge2YaRQ2XyolQL30EzTSo//z9SzANBgkqhkiG9w0BAQUF
AAOCAQEA1nPnfE920I2/7LqivjTFKDK1fPxsnCwrvQmeU79rXqoRSLblCKOzyj1hTdNGCbM+w6Dj
Y1Ub8rrvrTnhQ7k4o+YviiY776BQVvnGCv04zcQLcFGUl5gE38NflNUVyRRBnMRddWQVDf9VMOyG
j/8N7yy5Y0b2qvzfvGn9LhJIZJrglfCm7ymPAbEVtQwdpf5pLGkkeB6zpxxxYu7KyJesF12KwvhH
hm4qxFYxldBniYUr+WymXUadDKqC5JlR3XC321Y9YeRq4VzW9v493kHMB65jUr9TU/Qr6cf9tveC
X4XSQRjbgbMEHMUfpIBvFSDJ3gyICh3WZlXi/EjJKSZp4A==
-----END CERTIFICATE-----

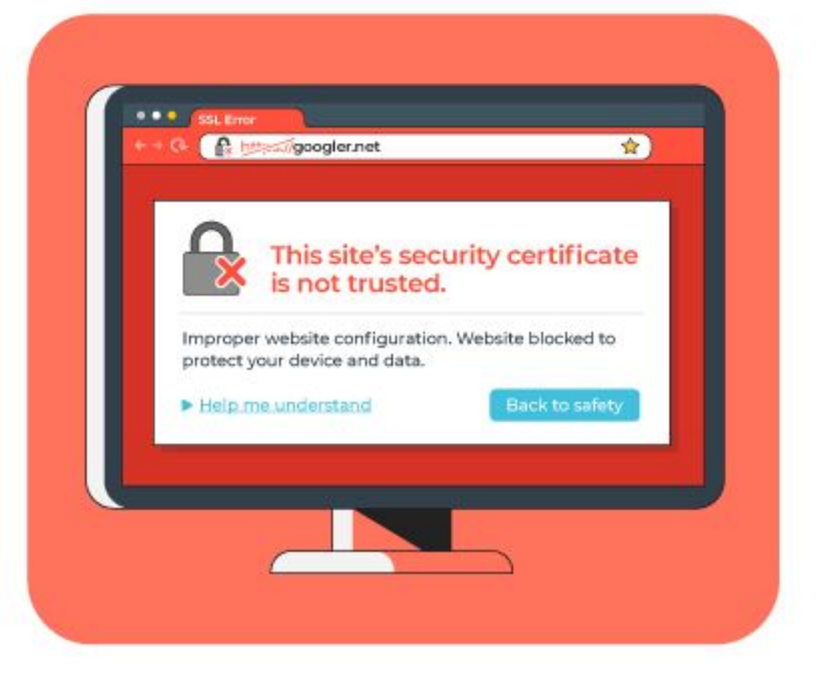
Entrust.net Premium 2048 Secure Server CA
=========================================
-----BEGIN CERTIFICATE-----
MIIEKjCCAxKgAwIBAgIEOGPe+DANBgkqhkiG9w0BAQUFADCBtDEUMBIGA1UEChMLRW50cnVzdC5u
ZXQxQDA+BgNVBAsUN3d3dy5lbnRydXN0Lm5ldC9DUFNfMjA0OCBpbmNvcnAuIGJ5IHJlZi4gKGxp
bWl0cyBsaWFiLikxJTAjBgNVBAsTHChjKSAxOTk5IEVudHJ1c3QubmV0IExpbWl0ZWQxMzAxBgNV
BAMTKkVudHJ1c3QubmV0IENlcnRpZmljYXRpb24gQXV0aG9yaXR5ICgyMDQ4KTAeFw05OTEyMjQx
NzUwNTFaFw0yOTA3MjQxNDE1MTJaMIG0MRQwEgYDVQQKEwtFbnRydXN0Lm5ldDFAMD4GA1UECxQ3
d3d3LmVudHJ1c3QubmV0L0NQU18yMDQ4IGluY29ycC4gYnkgcmVmLiAobGltaXRzIGxpYWIuKTEl
MCMGA1UECxMcKGMpIDE5OTkgRW50cnVzdC5uZXQgTGltaXRlZDEzMDEGA1UEAxMqRW50cnVzdC5u
ZXQgQ2VydGlmaWNhdGlvbiBBdXRob3JpdHkgKDIwNDgpMIIBIjANBgkqhkiG9w0BAQEFAAOCAQ8A
MIIBCgKCAQEArU1LqRKGsuqjIAcVFmQqK0vRvwtKTY7tgHalZ7d4QMBzQshowNtTK91euHaYNZOL
Gp18EzoOH1u3Hs/lJBQesYGpjX24zGtLA/ECDNyrpUAkAH90lKGdCCmziAv1h3edVc3kw37XamSr
hRSGlVuXMlBvPci6Zgzj/L24ScF2iUkZ/cCovYmjZy/Gn7xxGWC4LeksyZB2ZnuU4q941mVTXTzW
nLLPKQP5L6RQstRIzgUyVYr9smRMDuSYB3Xbf9+5CFVghTAp+XtIpGmG4zU/HoZdenoVve8AjhUi
VBcAkCaTvA5JaJG/+EfTnZVCwQ5N328mz8MYIWJmQ3DW1cAH4QIDAQABo0IwQDAOBgNVHQ8BAf8E
```

CERTIFICATE AUTHORITY



SELF-SIGNED CERTIFICATE





CA certificates extracted from Mozilla Firefox

https://curl.se/docs/caextract.html

http://curl.haxx.se/ca/cacert.pem

Date	Certificates
2022-10-11 (sha256)	142
2022-07-19 (sha256)	140
2022-04-26 (sha256)	135
2022-03-29 (sha256)	132
2022-03-18 (sha256)	132
2022-02-01 (sha256)	133
2021-10-26 (sha256)	130
2021-09-30 (sha256)	127
2021-07-05 (sha256)	128
2021-05-25 (sha256)	127

```php
<?php
$url = "https://....xml";

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
$fp = fopen("file.xml", "w");
curl_setopt($ch, CURLOPT_FILE, $fp);
curl_setopt($ch, CURLOPT_VERBOSE, true);

curl_setopt($ch, CURLOPT_CAINFO, 'cacert.pem');

if (curl_exec($ch) === false) {
    echo 'Ошибка curl: ' . curl_error($ch);
} else {
    echo 'Операция завершена без каких-либо ошибок';
}
curl_close($ch);
fclose($fp);
```

Загрузка: проблема https

```
* Trying 1.1.1.1:443…
* TCP_NODELAY set
* Connected to yet.another.shop (1.1.1.1) port 443 (#0)
* ALPN, offering h2
* ALPN, offering http/1.1
* successfully set certificate verify locations:
* CAfile: cacert.pem
   CApath: /etc/ssl/certs
* SSL certificate problem: unable to get local issuer
certificate
* Closing connection
```



Warning: Potential Security Risk Ahead

Firefox detected a potential security threat and did not continue to untrusted-root.badssl.com. If you visit this site, attackers could try to steal information like your passwords, emails, or credit card details.

What can you do about it?

The issue is most likely with the website, and there is nothing you can do to resolve it.

If you are on a corporate network or using anti-virus software, you can reach out to the support teams for assistance. You can also notify the website's administrator about the problem.

Learn more...

Go Back (Recommended) | Advanced...

Someone could be trying to impersonate the site and you should not continue.

Websites prove their identity via certificates. Firefox does not trust untrusted-root.badssl.com because its certificate issuer is unknown, the certificate is self-signed, or the server is not sending the correct intermediate certificates.

Error code: SEC_ERROR_UNKNOWN_ISSUER

View Certificate

Go Back (Recommended) | Accept the Risk and Continue

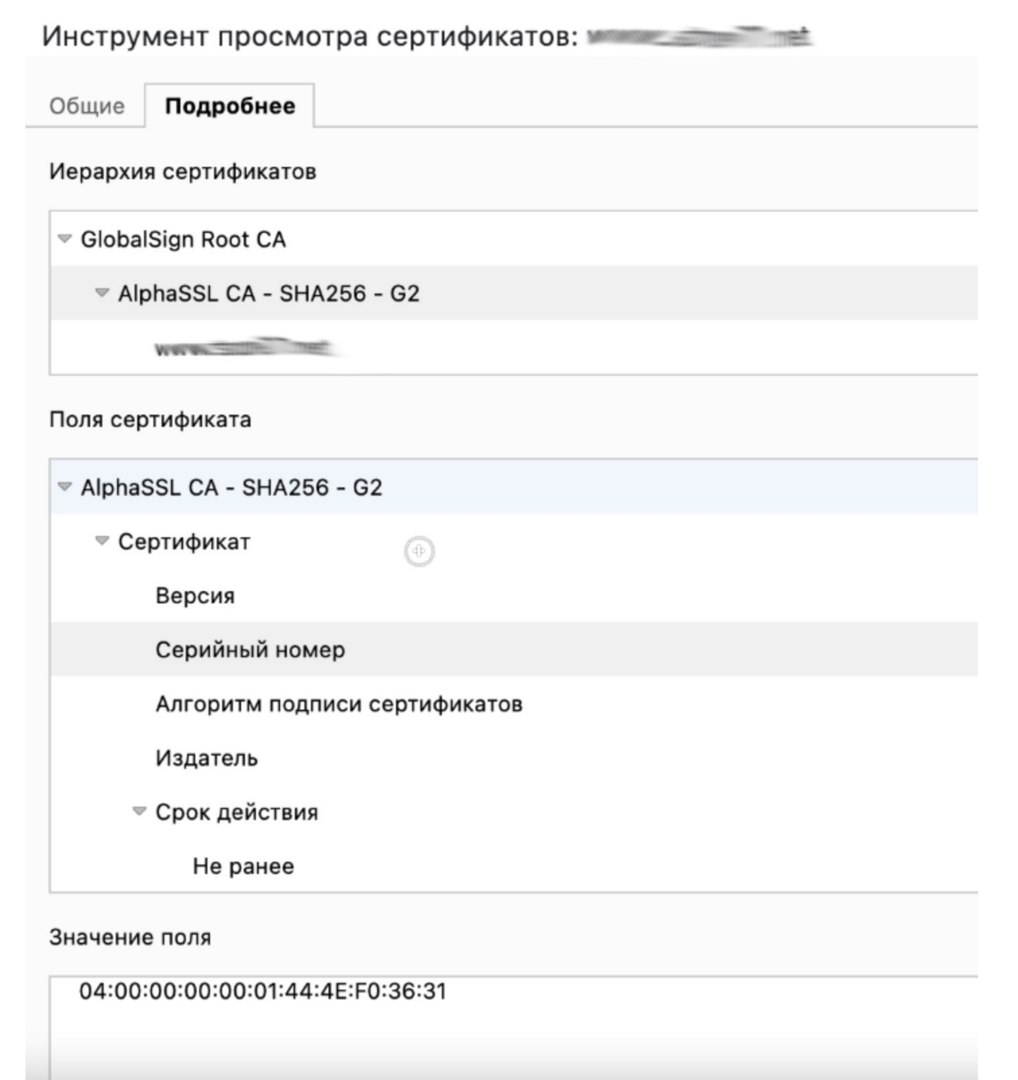
☐ Report errors like this to help Mozilla identify and block malicious sites

Корневые сертификаты

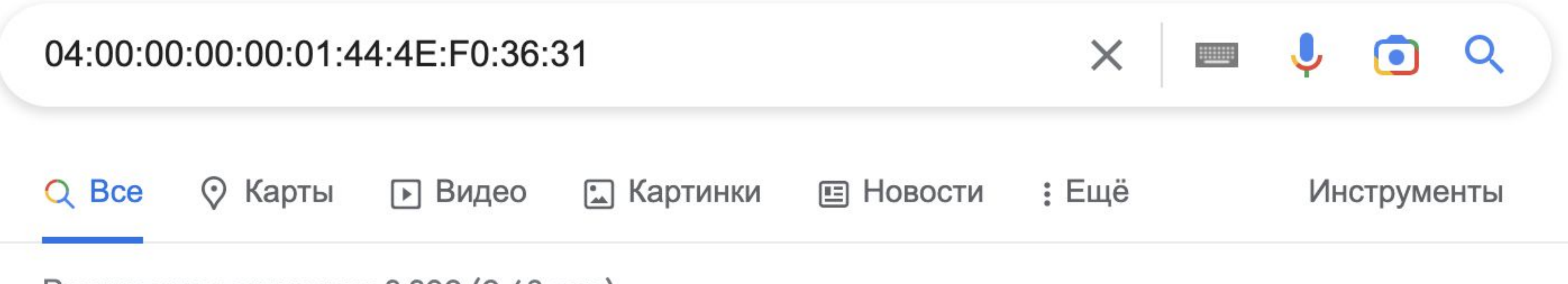
Mozilla
https://wiki.mozilla.org/CA/Included_Certificates

Chrome Root Store(soon)

https://chromium.googlesource.com/chromium/src/+/main/net/data/ssl/chrome_root_store/root_store.md


Инструмент просмотра сертификатов:
Общие
Подробнее
Иерархия сертификатов
GlobalSign Root CA
AlphaSSL CA - SHA256 - G2
Поля сертификата
AlphaSSL CA - SHA256 - G2
Сертификат
Версия
Серийный номер
Алгоритм подписи сертификатов
Издатель
Срок действия
Не ранее
Значение поля
04:00:00:00:00:01:44:4E:F0:36:31

Загрузка: https – AlphaSSL



Результатов: примерно 3 390 (0,68 сек.)

https://support.globalsign.com › ... ▾ Перевести эту страницу

AlphaSSL Intermediate Certificates - GlobalSign Support

Valid until: 20 February 2024. Serial #: **04 00 00 00 00 01 44 4e f0 36 31**. Thumbprint: 4c 27 43 17 17 56 5a 3a 07 f3 e6 d0 03 2c 42 58 94 9c f9 ec.

Загрузка: https – AlphaSSL

SHA-256 Orders (Default)

AlphaSSL SHA-256 R1 Intermediate Certificate

AlphaSSL CA - SHA256 - G2
SHA256 • RSA • 2048

Valid until: 20 February 2024
Serial #: 04 00 00 00 00 01 44 4e f0 36 31
Thumbprint: 4c 27 43 17 17 56 5a 3a 07 f3 e6 d0 03 2c 42 58 94 9c f9 ec

Download Certificate (Binary/DER Encoded) | View in Base64

Загрузка: https: успех

```
Connected to another.shop (1.1.1.1) port 443(#0)
 *ALPN: offers h2
 *ALPN: offers http/1.1
 *CAfile: cacert.pem
 *CApath: none
 *SSL connection using TLSv1.3 / TLS_AES_256_GCM_SHA384
 *ALPN: server accepted h2
 *Server certificate:
 *subject: CN=another.shop
 *start date: Mar 5 14:44:11 2022 GMT
 *expire date: Apr 14:44:10 2023 GMT
 *subjectAltName: host "another.shop" matched cert's "another.shop"
 *issuer: C=BE; O=GlobalSign nv-sa; CN=AlphaSSL CA - SHA256 - G2
 *SSL certificate verify ok.
```

Вывод 2

SSL – сложная штука.

В разных браузерах работает по-разному.

Нужно устранять неопределенность путем фиксирования и обновления списка корневых сертификатов.

Проблема с большим XML

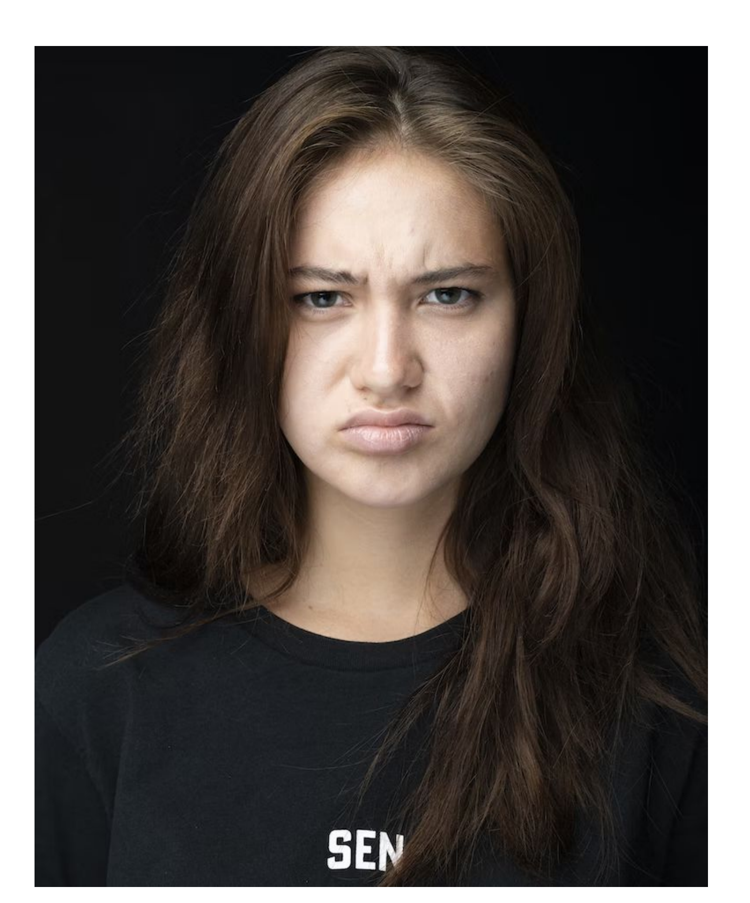
SUPPORT: Не грузится один из фидов!
В хроме вроде открывается, но страничка виснет

DEV: А большой фид?

SUPPORT: ~ **510 MB**

DEV:





Смотрим в логи

```
…
Out of Memory: Killed process
…
```

Проблема с большим XML- memory_get_usage

```php
<?php

$content= file_get_contents('510M_feed.xml');
$xml = simplexml_load_string($content);

echo "Memory usage:".round(memory_get_usage()/1024/1024, 2)."MB";

// Memory usage: 487.14 MB
```

Проблема с большим XML - memory_get_usage(true)

```
<?php

$content= file_get_contents('510M_feed.xml');
$xml = simplexml_load_string($content);

echo "Memory usage:".round(memory_get_usage(true)/1024/1024,2).
"MB";

// Memory usage: 488.77 MB
```

Проблема с большим XML

```
<?php

$xml = simplexml_load_file('510M_feed.xml');

echo "Memory usage: " .
round(memory_get_usage(true)/1024/1024, 2) . " MB";

// Memory usage: 2 MB
```

Проблема с большим XML

```
<?php
ini_set('memory_limit', '1024M');
$xml = simplexml_load_file('510MB_feed.xml');

echo "Memory usage: " .
round(memory_get_usage(true)/1024/1024, 2) . " MB";

// Memory usage: 2 MB
```

Проблема с большим XML

```
<?php
ini_set('memory_limit', '1M');
$xml = simplexml_load_file('510MB_feed.xml');

echo "Memory usage: " .
round(memory_get_usage(true)/1024/1024, 2) . " MB";

// Memory usage: 2 MB
```

Проблема с большим XML

```
<?php
ini_set('memory_limit', '1M');
$content= file_get_contents('510M_feed.xml');
$xml = simplexml_load_string($content);

echo "Memory usage:".round(memory_get_usage(true)/1024/1024,
2)."MB";

// PHP Fatal error:  Allowed memory size of 2097152 bytes
exhausted (tried to allocate 510419376 bytes)...
```

Проблема с большим XML

```
<?php

$xml = simplexml_load_file('510MB_feed.xml');
unlink('510MB_feed.xml');
foreach($xml as $el=>$v);

echo "Memory usage:".round(memory_get_usage(true)/1024/1024,
2)."MB";

// Memory usage: 2 MB
```

Проблема с большим XML – документация PHP

https://www.php.net/manual/en/function.memory-get-usage.php

memory_get_usage

▲1▼ ad-rotator.com 18 years ago

```
The method sandeepc at myrealbox dot com posted yields larger memory usage, my guess is that it includes all the PHP
interpreter/internal code and not just the script being run.

1) Use ps command
MEMORY USAGE (% KB PID ):  0.8 12588 25087 -> about 12MB
2) Use memory_get_usage()
int(6041952) -> about 6MB
```

Проблема с большим XML – расход памяти

```
ps aux --sort=-%mem
USER       PID %CPU %MEM     VSZ    RSS TTY      STAT START   TIME COMMAND
user       283  8.5 29.3 4861164 4809100 pts/2 S+   11:10   0:04 php load.php
```

VSZ – virtual memory size of the process in KiB (1024-byte units).

RSS – resident set size, the non-swapped physical memory that a task has used (in kilobytes)

PHP уже не торт!

RSS=4809100 KiB
RSS=4.8 GB !!!

memory_get_usage=2 MB



Измеряем память – PS- way

```php
<?php
function getUsagePs()
{
 $ps_output=exec("ps --pid ".getmypid()."--no-headers -o rss");
 return (int)$ps_output * 1024;
}
```

Проблема с большим XML – измеряем с PS

```php
<?php

simplexml_load_file('510mb_feed.xml');

echo round(memory_get_usage(true)/1024/1024, 2) . " MB";
echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2) . " MB";

// 2 MB
// PS memory: 4209.34 MB
```

Проблема с большим XML – измеряем с PS

```
echo round(memory_get_usage(true)/1024/1024, 2) . " MB";
echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2) . " MB";

// 2 MB
// PS memory: 24.98 MB

simplexml_load_file('510mb_feed.xml');

echo round(memory_get_usage(true)/1024/1024, 2) . " MB";
echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2) . " MB";

// 2 MB
// PS memory: 4209.34 MB
```

Проблема с большим XML – XmlReader

```
$reader = new XmlReader();
$reader->open('510mb_feed.xml');

while ($reader->read()) {

}

echo round(memory_get_usage(true)/1024/1024, 2) . " MB";
echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2) ."
MB";

// 2 MB
// PS memory: 25.39 MB
```

```php
$reader = new XmlReader();
$reader->open('510mb_feed.xml');
$i=0;

while ($reader->read()) {
   if (++$i%5000000==0) {
       echo "\n".$i;
       echo "\n".round(memory_get_usage(true)/1024/1024, 2)." MB";
       echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2)."MB";
   }
}

echo round(memory_get_usage(true)/1024/1024, 2) . " MB";
echo "\nPS memory: ".round(getUsagePs()/1024/1024, 2) . " MB";
```

Проблема с большим XML – проверяем XMLReader

```
5000000 2 MB; PS memory: 25.45 MB
10000000 2 MB; PS memory: 25.47 MB
15000000 2 MB; PS memory: 25.49 MB
20000000 2 MB; PS memory: 25.5 MB
25000000 2 MB; PS memory: 25.5 MB
```

Как избежать утечек?

Docker\k8s limits?

Тесты?



Добавляем контроль над расходом в тесты

```
$this->assertLessThan(100000, memory_get_peak_usage(true),
‘Memory usage too big, possible memory leak’);
```

```php
class Memory
{
    private static int $maxMemory = 0;

    public function getUsage($updateMaxMemory = true): int
    {
        $memory = $this->getUsagePs();
        if ($updateMaxMemory) {
            self::$maxMemory = max(self::$maxMemory, $memory);
        }
        return $memory;
    }

    public function getPeakUsage()...
    public function reset()...
    public function getUsagePs()..
}
```

Контролируем память в тестах

```
$loader = new SourceLoader($http);

Memory::reset();

$loader->load($feed);
$this->assertLessThan(123,Memory::getPeakUsage(),
                                   'Potential memory leak!')
```

Вывод 3

PHP не “видит” всю память, которую использует и **memory_get_usage**, **memory_get_peak_usage**, **memory_limit** доверять нельзя.

Проблема – медленная вставка в ElasticSearch

CXO: Данные в поиске старые!

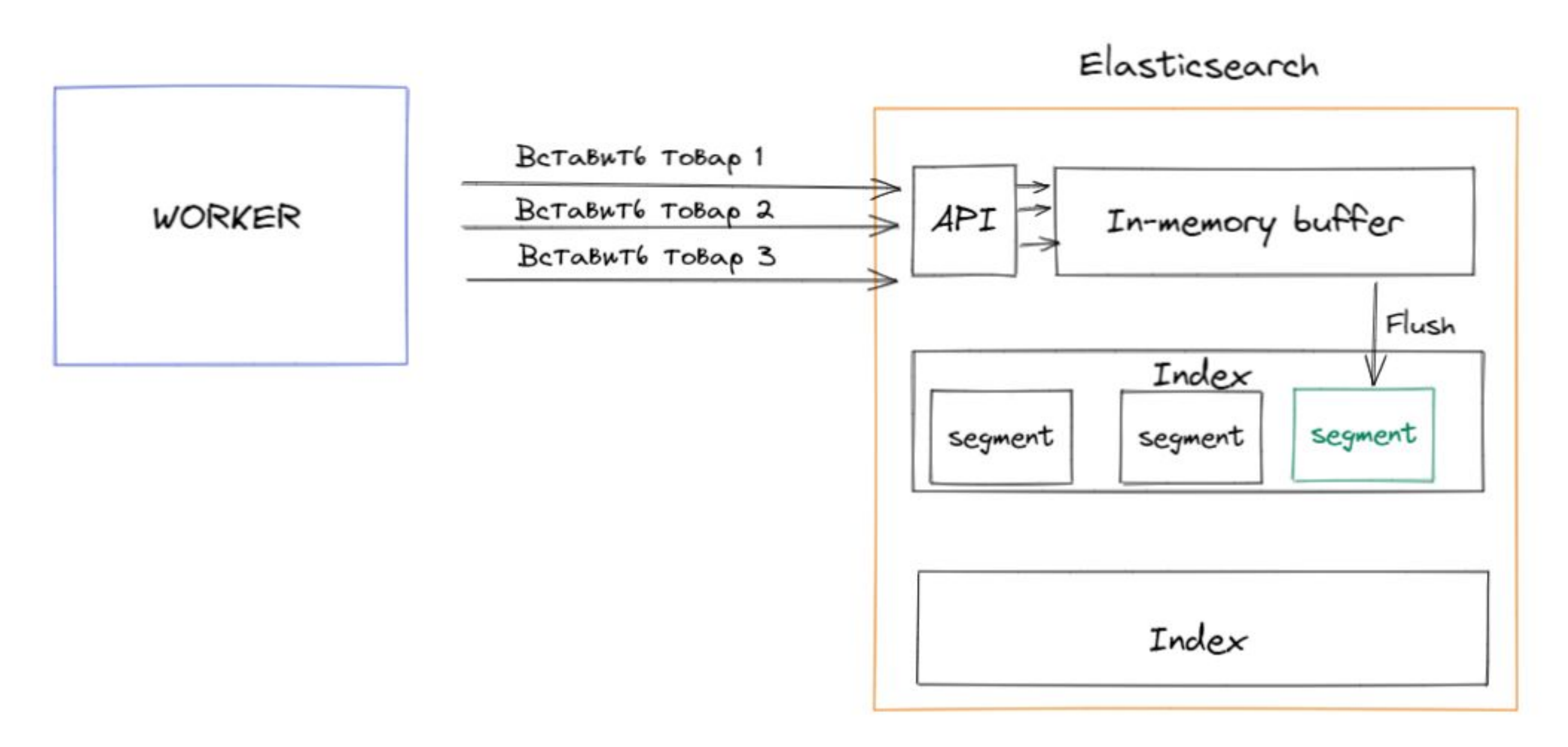
DEV: А в фидах новые?

CXO: В фидах новые, а в поиске старые!

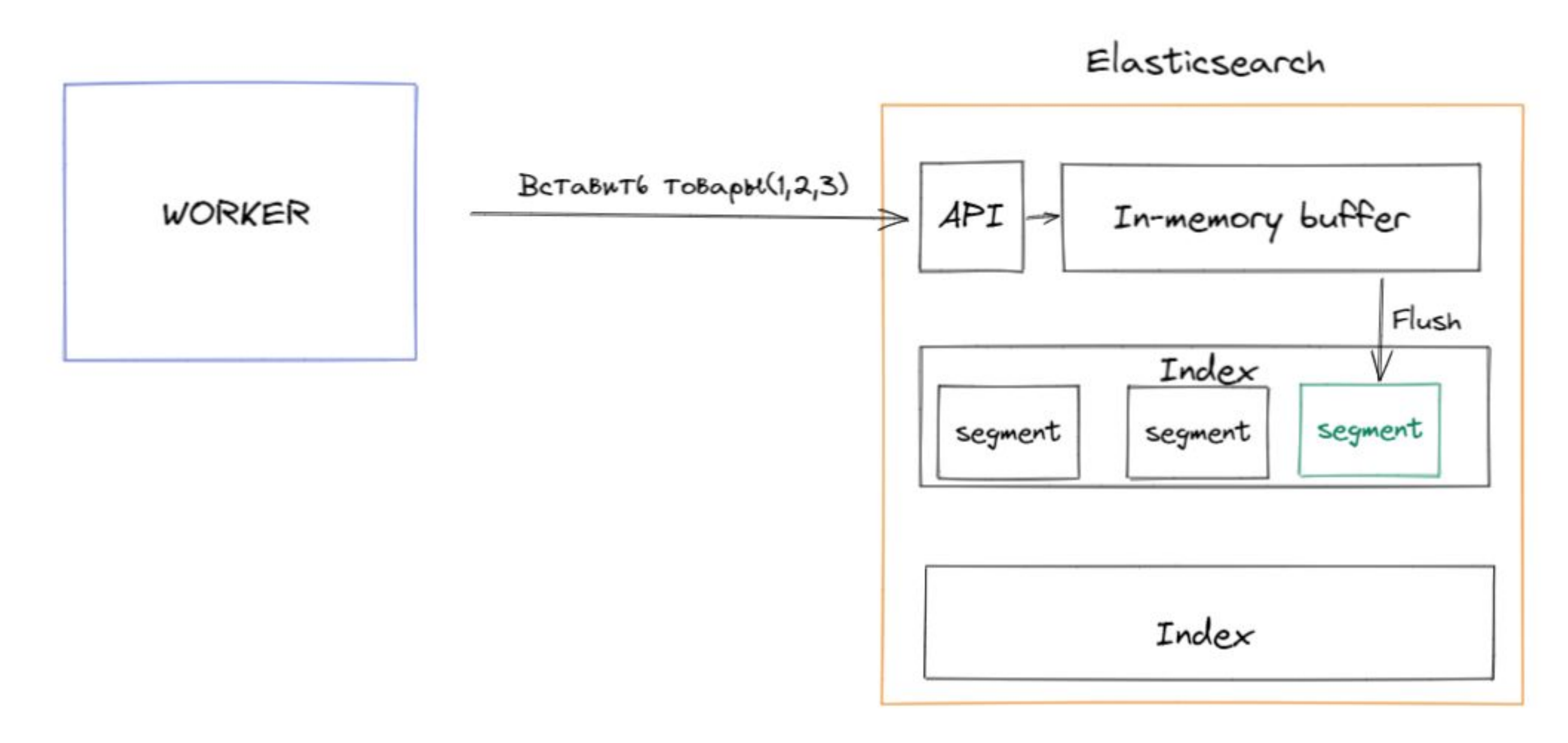



elastic
1-55181
2-44077

“Секрет” 1 – Батчинг

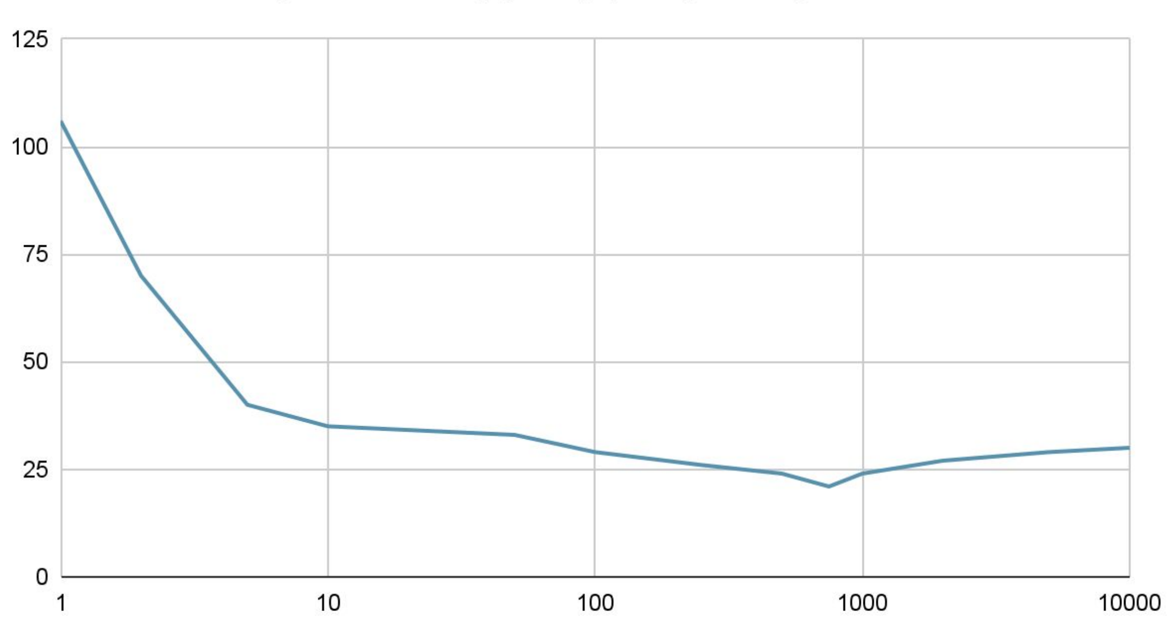


“Секрет” 1 – Батчинг

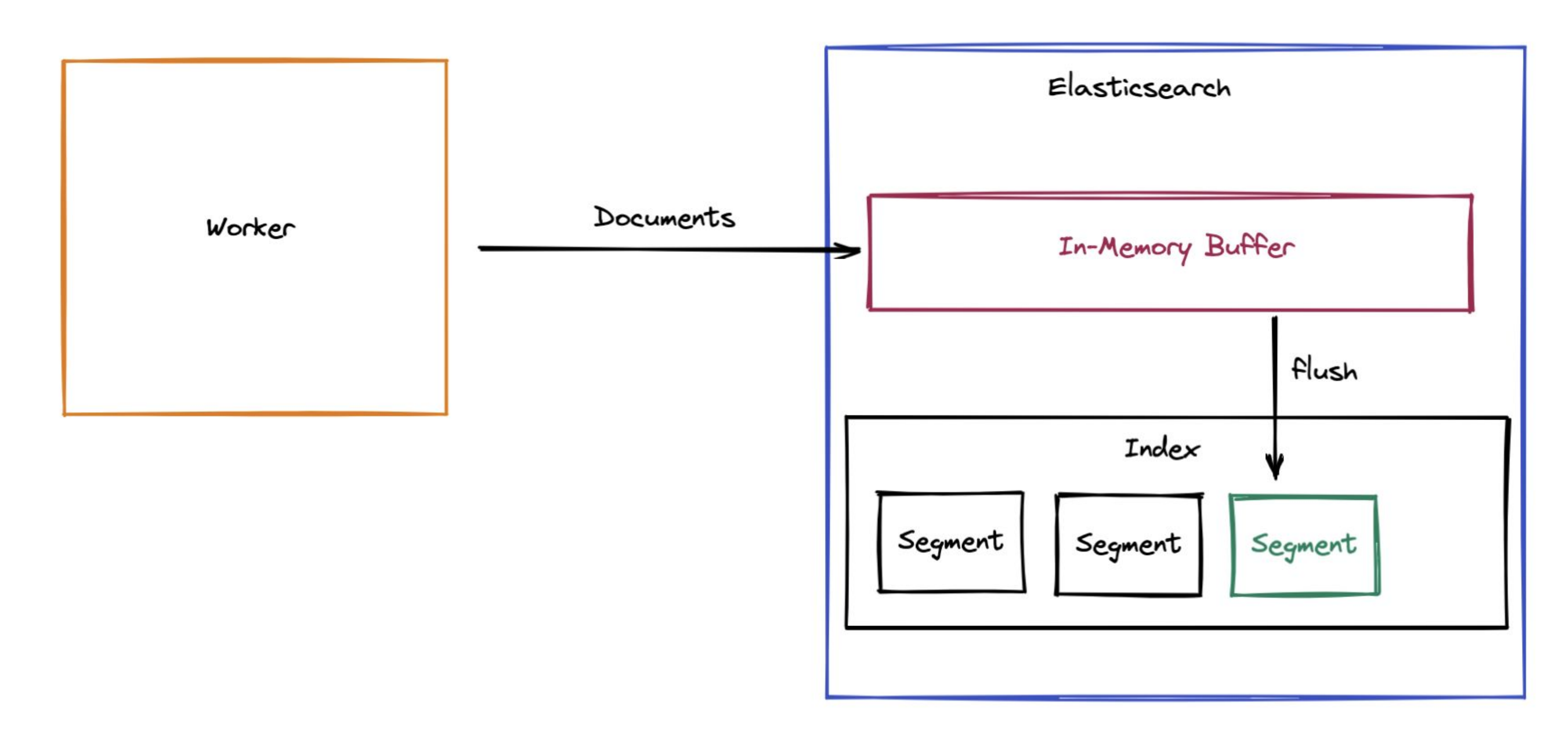


“Секрет” 1 – Батчинг

Время загрузки тестового фида на 100 000 товаров



“Секрет” 2 – refresh_interval



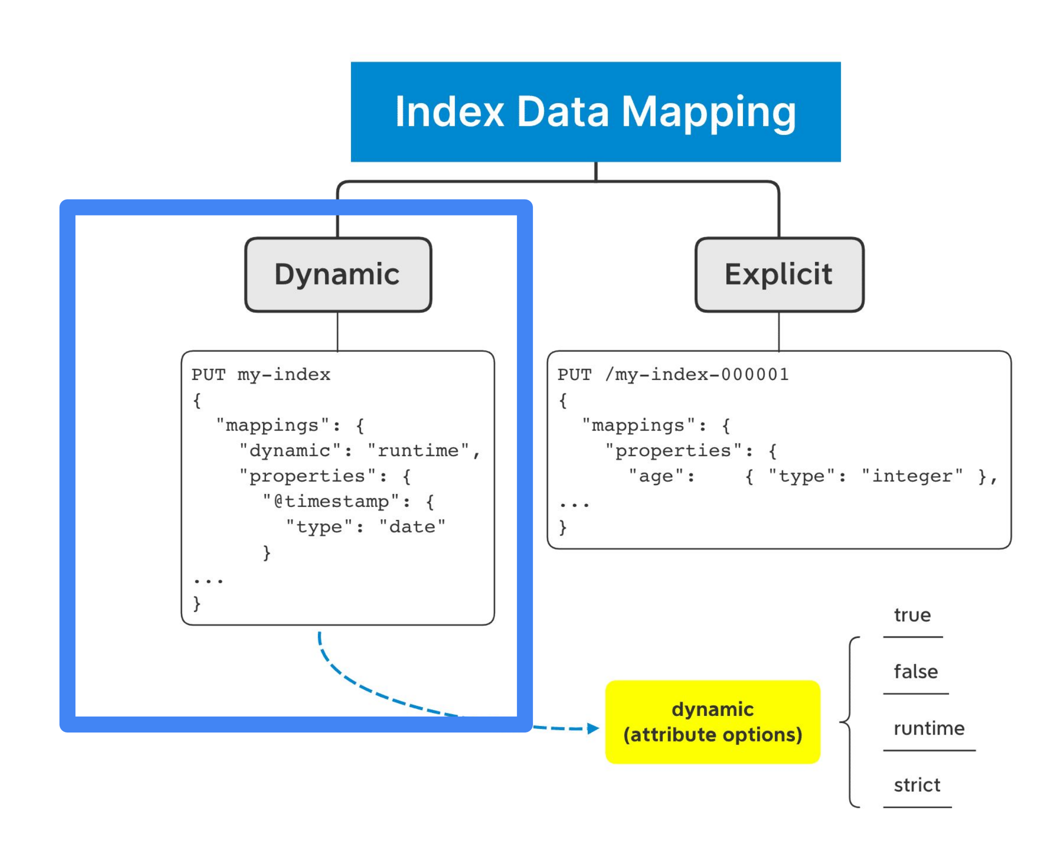
“Секрет” 2 – refresh_interval

Индексируем только после полной вставки фида в Elasticsearch

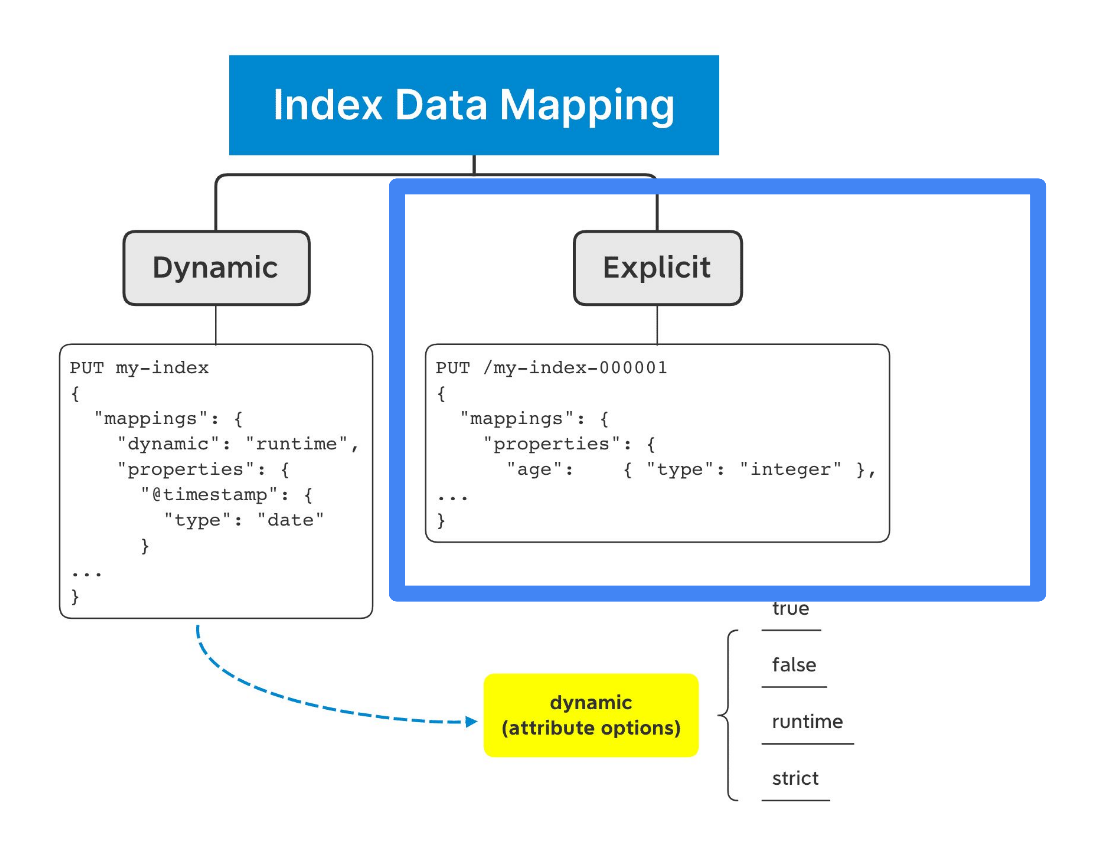
1. Создаем новый индекс с refresh_interval=-1
2. Вставляем ВСЕ документы
3. Force merge

Ускорение
41 сек => 22 сек   1.86X!

Маппинг типов данных: динамический



Маппинг типа полей: явный



Секрет 3 – Фиксированный маппинг типа полей

1) Пробегаемся по всему фиду, определяем типы полей (число, строка)
2) Создаем индекс с необходимыми полями
3) Индексируем фид

Ускорение
150 сек => 22 сек 6.81X!

Определение типов полей данных

Товар	Вес	Тип поля Вес	Диагональ	Тип поля Диагональ
1	1000	число	22	число

Определение типов полей данных

Товар	Вес	Тип поля Вес	Диагональ	Тип поля Диагональ
1	1000	число	22	число
2	332	число	25 дюймов	строка

Определение типов полей данных

Товар	Вес	Тип поля Вес	Диагональ	Тип поля Диагональ
1	1000	число	22	число
2	332	число	25”	строка
3	533	**число**	33	**строка**

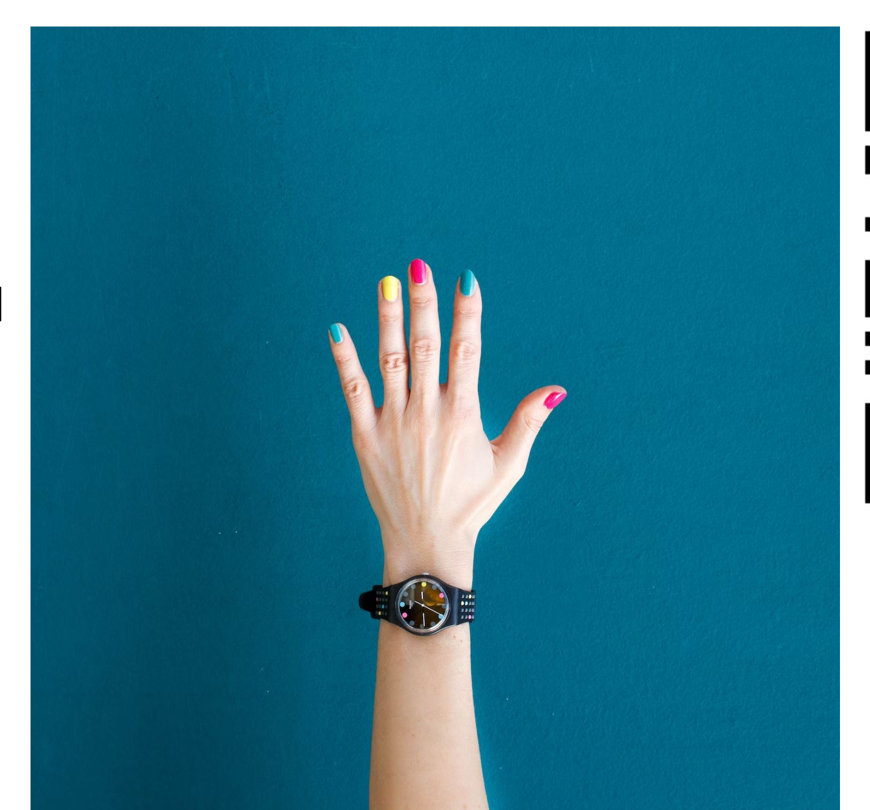
Вывод 4

Для ускорения вставки данных в БД следует использовать как универсальные подходы (батчинг), так и тонкую настройку специализированных параметров, которые могут все ускорить в разы.

Вопросы?

Красников Иван
searchbooster.io

telegram: king5551





ОЦЕНИТЬ ДОКЛАД